The Journal of Chemical Physics

Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability

Cite as: J. Chem. Phys. **156**, 084104 (2022); https://doi.org/10.1063/5.0079574
Submitted: 22 November 2021 • Accepted: 31 January 2022 • Published Online: 22 February 2022

Thijs Stuyver and Connor W. Coley

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability

Cite as: J. Chem. Phys. 156, 084104 (2022); doi: 10.1063/5.0079574
Submitted: 22 November 2021 • Accepted: 31 January 2022 •
Published Online: 22 February 2022

Thijs Stuyver[1] and Connor W. Coley[1,2,a)]

AFFILIATIONS

[1] Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA
[2] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA

Note: This paper is part of the JCP Special Topic on Chemical Design by Artificial Intelligence.
[a)]**Author to whom correspondence should be addressed:** ccoley@mit.edu

ABSTRACT

There is a perceived dichotomy between structure-based and descriptor-based molecular representations used for predictive chemistry tasks. Here, we study the performance, generalizability, and explainability of the quantum mechanics-augmented graph neural network (ml-QM-GNN) architecture as applied to the prediction of regioselectivity (classification) and of activation energies (regression). In our hybrid QM-augmented model architecture, structure-based representations are first used to predict a set of atom- and bond-level reactivity descriptors derived from density functional theory calculations. These estimated reactivity descriptors are combined with the original structure-based representation to make the final reactivity prediction. We demonstrate that our model architecture leads to significant improvements over structure-based GNNs in not only overall accuracy but also in generalization to unseen compounds. Even when provided training sets of only a couple hundred labeled data points, the ml-QM-GNN outperforms other state-of-the-art structure-based architectures that have been applied to these tasks as well as descriptor-based (linear) regressions. As a primary contribution of this work, we demonstrate a bridge between data-driven predictions and conceptual frameworks commonly used to gain qualitative insights into reactivity phenomena, taking advantage of the fact that our models are grounded in (but not restricted to) QM descriptors. This effort results in a productive synergy between theory and data science, wherein QM-augmented models provide a data-driven confirmation of previous qualitative analyses, and these analyses in turn facilitate insights into the decision-making process occurring within ml-QM-GNNs.

I. INTRODUCTION

The rationalization and prediction of reactivity trends is one of the core objectives of (theoretical) chemistry. Before advanced *ab initio* quantum mechanics (QM) computations became a commonplace, a plethora of heuristic concepts and qualitative theory-inspired rules—usually tailored to a specific subclass of compounds or reactions—were already proposed and developed to this end. Some iconic examples in this regard are the Woodward–Hoffmann rules for the prediction of reaction outcomes of pericyclic transformations,[1] the Bell–Evans–Polanyi principle,[2] and the hard–soft acid–base (HSAB) concept pioneered by Pearson.[3]

Later on, several competing, overarching theoretical frameworks emerged, all of which aspire to describe chemical reactions in a universally applicable manner. Some well-known examples of such frameworks are conceptual density functional theory (cDFT),[4,5] the (molecular orbital-based) activation-strain model (ASM),[6,7] and the valence bond (VB) reactivity model.[8,9] At the core of each of these frameworks is the definition of a limited set of chemically meaningful quantities or descriptors. These probe, either directly or indirectly, the magnitude of different fundamental interactions (e.g., orbital-based or "soft–soft" vs electrostatic or "hard–hard" interactions in cDFT), which collectively characterize the chemical system and enable an insightful and internally consistent

(qualitative) discussion of its reactivity. As these frameworks matured, a concerted effort has taken place to incorporate/embed the various previously proposed concepts and rules, albeit of limited scope, into them.[10,11] Attempts have also been made to build bridges between the individual frameworks themselves, thus facilitating enhanced understanding.[12–15]

With the advent of machine learning (ML) and artificial intelligence methods, an entirely different approach to chemical reactivity prediction emerged, not constrained by any specific theoretical framework. Recent work in this area includes the prediction of reaction products,[16–18] reaction yields,[19–21] and bond dissociation energies.[22,23] Instead of building physically motivated representations of the constituent molecules (e.g., of reactants, products, and catalysts), many ML approaches take advantage of a model's ability to learn meaningful representations and start with simple structural descriptors/features. These include graph-based molecular representations using graph neural networks (GNNs),[22–25] simpler molecular fingerprint[26] representations,[16,27] and even simplified molecular-input line-entry system (SMILES) representations.[21,28,29] Such structure-based representations do not necessarily have a *direct* connection to reactivity and rely on the nonlinearity and expressivity of ML models to relate the structure to function.[30] Nevertheless, given sufficient data to train these models, deep learning for reactivity prediction has been demonstrated to achieve accurate predictions of reaction products and/or energies.

Unfortunately, the setup of regular GNNs renders the decision-making process occurring inside them rather opaque. Rationalizing the predictions made by these networks, i.e., rendering them "explainable,"[31] has largely been limited to brittle techniques for estimating the sensitivity of predictions to atom- and bond-level contributions.[20,32–35] Therefore, these models have been often characterized as black boxes. Furthermore, structure-based GNNs tend to/are assumed to underperform in data-limited settings as they must learn a meaningful representation "from scratch."[36,37] This represents a significant drawback of GNNs, as large datasets with thousands of data points are rare in the field of chemistry, especially when focusing on subtle reactivity questions.[38] Other machine learning methods, e.g., kernel ridge regression (KRR)[39,40] or Gaussian process (GP) regression[37] in combination with molecular representations such as bag-of-bonds (BoB),[41] FCHL19,[42] or spectrum of London and Axilrod–Teller–Muto (SLATM) potentials[43] are generally assumed to perform better in the face of data scarcity; even then, several thousand data points may be needed to train a model to an acceptable level of accuracy.[37,39,40]

One strategy to mitigate the drawbacks of these data-hungry methods is to represent molecules with functional descriptors that have a more direct (and linear) relationship with their reactivity. Instead of working within a specific theoretical framework, reactivity is modeled through statistical methods, but descriptors/features used as input are inspired by the chemistry and physics underlying the investigated reactivity problem. Champions of this approach are Sigman *et al.*, among others, who have employed multivariate linear regression to relate sophisticated electronic and steric descriptors to complex properties such as enantioselectivity,[44] and Zahrt *et al.*, who used support vector machines (SVMs) and feed-forward neural networks for the same task.[45] Recently, Ahneman *et al.* demonstrated the prediction of reaction yields of C–N cross-coupling reactions with a random forest model by selecting reaction-specific descriptors, starting from ~4000 data points obtained via high-throughput experimentation.[20] Other recent examples of this strategy can be found in the work of Beker *et al.*,[46] Li *et al.*,[47] and Jorner *et al.*[48]

While descriptor-based methods may require less data, be more generalizable, i.e., achieve an improved performance on classes of compounds not present in the training set,[46] and enable at least some explainability/interpretability compared to the use of universal/non-specific representations,[49] the selection of suitable descriptors is a non-trivial and problem-specific matter.[38] Even more importantly, in order to obtain these chemically relevant descriptors, a dataset specific computational workflow is often required, which creates a bottleneck that significantly hampers the ease of employability of this strategy.[38]

In this work, we build upon a unifying approach proposed by Guan *et al.*,[50] which aims to combine the advantages of structure-based GNNs with those of models based on expert-guided descriptors. Instead of feeding computationally expensive, task-specific descriptors directly to a machine learning model, we start from graph-based/structural input features but predict, as an intermediate step, a set of atom- and bond-level QM descriptors prior to the final reactivity prediction. Taking this approach, it becomes possible to construct a QM-based representation on-the-fly at minimal cost.

We assess the model performance on computational and experimental datasets (focusing on competing E2 vs S_N2 reactions[51] and aromatic substitution reactions,[50] respectively) and find that the resulting model architecture exhibits excellent accuracy in a data-limited regime. We further evaluate the model's ability to generalize to structures not seen during training using non-random data splits and observe a comparable improvement in performance. Most importantly, since our ml-QM-GNN models base their final reactivity predictions on a representation partially comprising QM descriptors, we are able to build a bridge to the traditional theoretical reactivity frameworks and explain the neural network's decision-making process in terms of traditional chemistry concepts. Overall, our work underscores that machine learning techniques and conceptual models are not mutually exclusive approaches but are able to benefit each other in a synergistic manner.

II. COMPUTATIONAL METHODS

A schematic overview of the complete QM-augmented neural network architecture used in this work is presented in Fig. 1 (cf. Sec. S1 of the supplementary material for an in-depth discussion of the individual network branches).

First, the simplified molecular-input line-entry system (SMILES) representations of compounds involved in the reaction are parsed into graph-based representations using RDKit,[52] and structural descriptors (atomic number, formal charge, ring status, bond order, etc.) are calculated for each heavy atom and bond. The structure-based representation is then used as input for a multitask GNN for QM descriptor prediction, based on a directed message passing neural network (D-MPNN) encoder, which has been adopted without modification from Guan *et al.*[50] (details related to the setup and training of this D-MPNN encoder are included in Sec. S1). Note that this multitask GNN part of our network shares several characteristics with the stand-alone DeepMoleNet

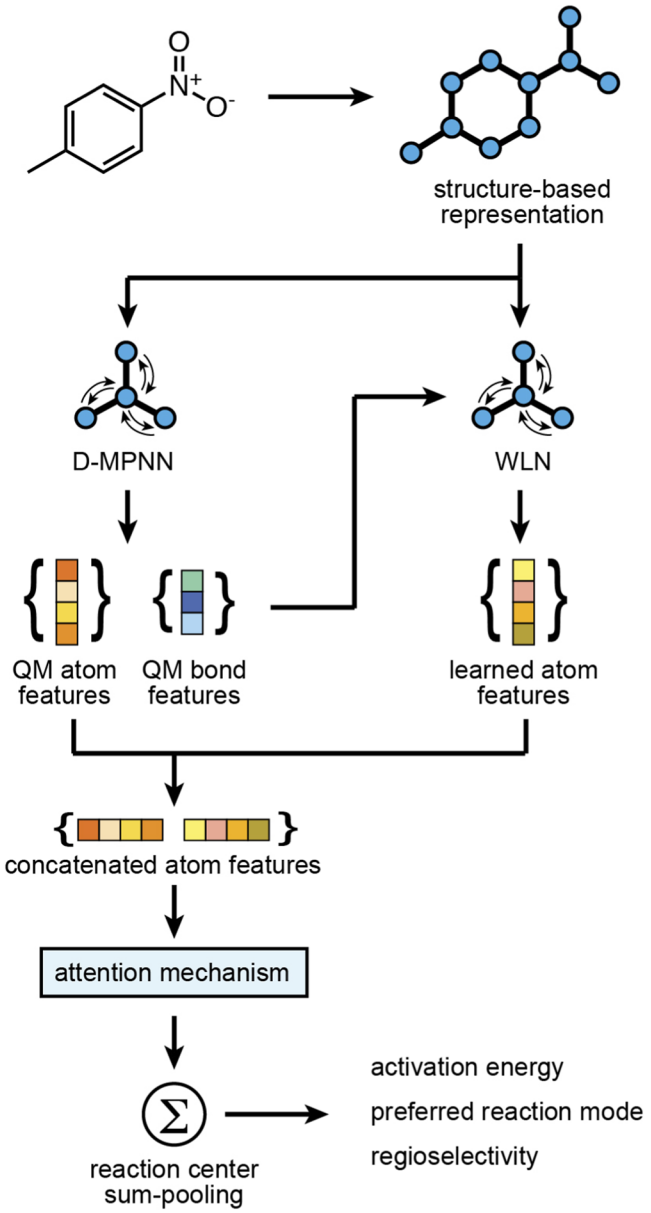

FIG. 1. Schematic overview of the ml-QM-GNN model architecture. WLN denotes the Weisfeiler–Lehman network branch, and D-MPNN denotes the directed message-passing branch of the network.

package and AIMNet, which have been developed to provide QM descriptors on-the-fly for qualitative reactivity analyses.[53,54]

The set of descriptors predicted by the D-MPNN encoder can be subdivided in two main categories: bond descriptors, i.e., natural population analysis (NPA) bond orders[55] and bond lengths, and atom-centered descriptors, i.e., (Hirshfeld) atomic charges,[56] as well as (atom-condensed) nucleo- and electrophilic Fukui functions,[5] and nuclear magnetic resonance (NMR) shielding constants.[57] In the finite-difference approximation,[58] the electrophilic Fukui function on site i (f_i^+) is defined as

$$f_i^+ = q_i(N) - q_i(N+1), \tag{1}$$

where $q_i(N)$ and $q_i(N+1)$ are the (Hirshfeld) partial charges on atom i for the corresponding (N)- and (N + 1)-electron wave function evaluated in the optimized N-electron geometry. The nucleophilic Fukui function (f_i^-) in its turn is defined as

$$f_i^- = q_i(N-1) - q_i(N), \tag{2}$$

where $q_i(N)$ and $q_i(N-1)$ are the (Hirshfeld) partial charges on atom i for the corresponding (N)- and (N − 1)-electron wave function evaluated in the optimized N-electron geometry.

The bond descriptors are converted to a vector representation through the application of a radial basis function (RBF) expansion and relayed to a separate Weisfeiler–Lehman network (WLN) branch, which combines them with the initial, structure-based representation in a convolutional embedding.[59] The atom-centered descriptors are kept separately and are similarly post-processed, i.e., they are scaled and then turned into a vector representation through RBF expansion.

The learned representation emerging from the WLN and the expanded atomic QM descriptor representation emerging from the D-MPNN branch of the network are subsequently concatenated, after which the concatenated representation is passed through a dense activation layer followed by a global attention mechanism[60] to capture the influence of distant parts of the reacting system. Because we focus on molecule- or reaction-level prediction tasks in this work, we aggregate these representations into one final feature vector by sum-pooling over the (hypothetically) reacting atoms. This global feature vector is transformed once more in a single-layer network to produce the barrier heights/activation energies in the case of a regression task and a regioselective preference in the case of a classification task (*vide infra*).

The regular GNN, used as the primary baseline in the discussions below, follows an analogous architecture as the ml-QM-GNN, except that it does not contain the QM descriptor network branch, i.e., the WLN operates on the original structural features alone and the attention mechanism takes only the learned features as input (cf. Sec. S1 of the supplementary material).

A. Datasets

We focus the evaluation of our ml-QM-GNN model on two publicly available datasets: one computational and one experimental (summary statistics, characterizing the distribution of the individual sets, are included in Sec. S3). The first dataset comprises computed stationary points along the potential energy surface for competing E2 and S_N2 reactions in the gas-phase, recently published by von Rudorff *et al.*[51] In total, four distinct nucleophiles (H^-, F^-, Cl^-, and Br^-), three distinct leaving groups (F^-, Cl^-, and Br^-), and permutations of five potential substituents (H, NO_2, CN, CH_3, and NH_2), on an ethyl-based scaffold, were considered [Fig. 2(a)]. As indicated by the authors, the substituents were selected to (i) maximize electronic effects, and (ii) minimize steric hindrance, which makes this dataset ideally suited for our ml-QM-GNN model based on electronic QM descriptors.

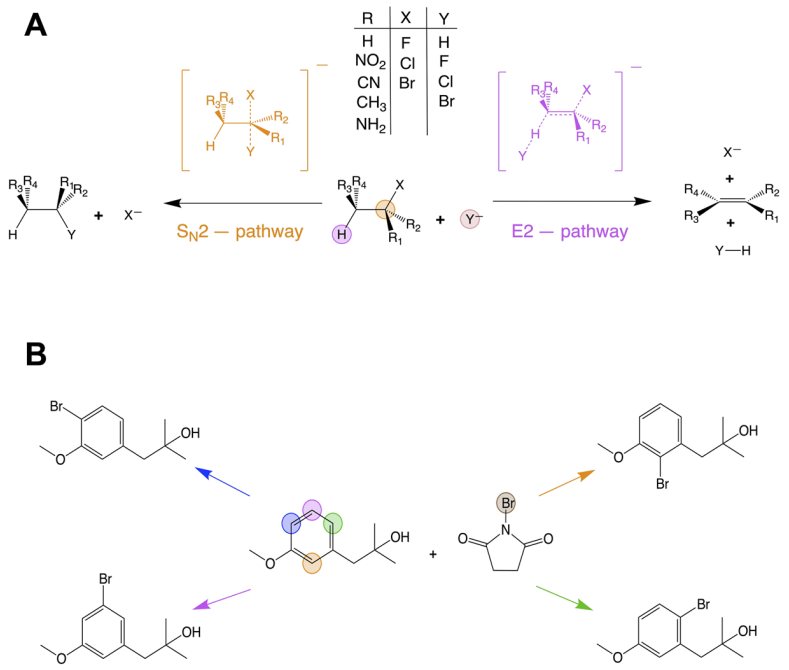

FIG. 2. Schematic representation of the two considered datasets. (a) The competing E2/S_N2 reaction pathways, with the respective attacking positions of the nucleophile indicated in purple/orange, respectively. The top center table gives an overview of the different substituents (R), leaving groups (X), and nucleophiles (Y) present in the dataset. (b) An example of a data point in the aromatic substitution reaction dataset. The colored dots indicate the potential reacting sites.

To construct a regression task, 3647 barrier heights calculated at the DF-LCCSD/cc-pVTZ//MP2/6-311G(d) level-of-theory[61–67] were extracted (1286 corresponding to E2 pathways; 2361 to S_N2).[40] An appropriate input for our GNNs was constructed by converting the 3D reactant complex geometry for each reaction into a 2D SMILES representation using xyz2mol.[68] Additional details about the data pre-processing can be found in Sec. S4.

To construct a classification task, we extracted both an E2 and S_N2 transition state (TS) from 791 unique reaction systems (Sec. S5). These data were used for training classification models to predict whether the E2 or S_N2 pathway is kinetically favored.

The second dataset we examine is a purely experimental one consisting of regioselective aromatic C–X substitution and C–H functionalization reactions from the Pistachio database, to which plausible side products corresponding to regiochemical alternatives, identified through template extraction and application,[69] were added.[70] This dataset was originally curated by Guan *et al.*[50] Since the data in Pistachio are only available to license-holders, Guan *et al.* filtered out a subset of 3242 data points for which the reactions are also present in the USPTO public database. Here, we examine the multi-way classification performance when predicting the regiochemical preference for this subset of reactions [Fig. 2(b)]. In other words, the model is trained to assign the highest probability to the correct/"true" product among the various regiochemical alternatives provided as input for each data point.

As we are mainly interested in understanding the effect of QM-augmentation under data-scarce conditions, models applied to this filtered dataset were primarily trained on a sample of only 200 data points. In each fold and iteration considered, these data points were randomly selected from the original, full training set. The small standard deviations in accuracy observed across different iterations (*vide infra*) suggest that the model is rather insensitive to the identity of the specific reactions sampled. Learning curves, visualizing how the accuracy of the model evolves as more data points are included during training, can be found in Sec. S6 of the supplementary material.

III. RESULTS AND DISCUSSION

A. Accuracy—E2/S_N2 dataset

As a first step, the performance of our ml-QM-GNN model was assessed by comparing the accuracy obtained for the barrier height/activation energy prediction for the competing E2/S_N2 reactions with the (structure-based) GNN baseline model. Initially, we split up the data according to the respective reaction type and performed three 5-fold cross-validations (CV) on each. In every iteration, the number of labeled data points considered for the construction of the model, i.e., the combination of training and validation set, was limited to 125, 250, 500, and 1000 in the case of E2, and 225, 450, 900, and 1800 points in the case of S_N2, matching the prior evaluation by the work of Heinen, von Rudorff, and von Lilienfeld.[40] The average mean absolute error (MAE) for our WL-based GNN models, obtained in this manner, are presented in the first two columns of Tables I and II.

The comparison between the MAEs obtained for the QM-augmented model and for the baseline model reveals that the inclusion of the QM descriptors in the model architecture results in a 5–6 kcal/mol lower error. The rate at which the accuracy improves as more labeled data points are included during training/validation is also improved.

Next to the WL-based GNNs, we also considered the performance of a powerful alternative structure-based neural network

TABLE I. Average MAE (kcal/mol) when predicting S_N2 barrier heights, obtained after three 5-fold CVs, for the baseline GNN, our ml-QM-GNN and Chemprop for different numbers of labeled data points. The standard deviations were determined based on the MAEs for the three replicates. The corresponding accuracies obtained from fivefold CV for the KRR models combined with BoB, SLATM, FCHL19, and one-hot encoding representations are included as well.[40] 20% of labeled points were reserved as a validation set for early stopping while training the GNN models.

Labeled points	Baseline GNN	ml-QM-GNN	Chemprop	BoB[a]	SLATM[a]	FCHL19[a]	One-hot encoding[a]	Multivariate regression
225 (180 + 45)	9.07 ± 0.04	3.61 ± 0.14	6.71 ± 0.08	4.89	4.44	3.80	3.53	6.43
450 (360 + 90)	8.89 ± 0.13	3.28 ± 0.04	4.01 ± 0.02	4.28	3.87	3.43	2.80	6.33
900 (720 + 180)	8.61 ± 0.07	2.97 ± 0.03	3.23 ± 0.07	3.78	3.21	3.11	2.42	6.40
1800 (1440 + 360)	8.49 ± 0.03	2.76 ± 0.01	2.85 ± 0.02	3.49	2.92	2.87	2.14	6.43

[a]Taken directly from the work of Heinen, von Rudorff, and von Lilienfeld.[40]

TABLE II. Comparison of the average MAE (kcal/mol) values on the predicted E2 barrier heights, obtained after three 5-fold CVs, for the baseline GNN, our ml-QM-GNN, and Chemprop for different numbers of labeled data points. Standard deviations were determined based on the MAEs for the three replicates. The corresponding accuracies obtained from fivefold CV for the KRR models combined with BoB, SLATM, FCHL19, and one-hot encoding representations are included as well. 20% of labeled points were reserved as a validation set for early stopping while training the GNN models.

Labeled points	Baseline GNN	ml-QM-GNN	Chemprop	BoB[a]	SLATM[a]	FCHL19[a]	One-hot encoding[a]	Multivariate regression
125 (100 + 25)	9.03 ± 0.18	4.08 ± 0.06	7.49 ± 0.10	4.67	4.43	4.01	3.53	6.43
250 (200 + 50)	8.78 ± 0.35	3.24 ± 0.08	5.78 ± 0.18	4.07	3.87	3.42	3.12	6.27
500 (400 + 100)	8.18 ± 0.04	2.91 ± 0.02	3.29 ± 0.09	3.71	3.21	3.01	2.69	5.90
1000 (800 + 200)	8.04 ± 0.17	2.65 ± 0.02	2.75 ± 0.06	3.27	2.92	2.75	2.40	6.07

[a]Taken directly from the work of Heinen, von Rudorff, and von Lilienfeld.[40]

architecture, Chemprop,[71] which has recently been developed by Yang *et al.*, for the same data splits, cf. the third column in Tables I and II as well as Sec. S7. Remarkably, Chemprop, which makes use of a D-MPNN[72] encoder for its convolutional embedding instead of a WL one, is significantly more data efficient than our baseline GNN for the considered task: Whereas for the smallest numbers of labeled data points, Chemprop does only marginally, i.e., 1–2 kcal/mol, better than the regular WL-based GNN, the advantage of the former grows rapidly as the number of labeled data points considered during training increases. Nevertheless, our ml-QM-GNN model outperforms Chemprop across the entire range of training/validation-set sizes considered. For the lowest training/validation-set sizes, the difference in accuracy is stark; the MAE obtained for the QM-augmented, WL-based GNN is almost half the MAE obtained for Chemprop. When a 1000 or more data points are considered during training, the advantage of our ml-QM-GNN dwindles to about 0.1 kcal/mol.

To contextualize the performance of our models, we also include kernel ridge regression (KRR) models benchmarked by Heinen *et al.* for four different (global) representations—BoB, SLATM, FCHL19, and one-hot encoding, respectively (in which every substituent site, as well as the nucleophile and leaving group site, is assigned a bit vector spanning all the possible substituent/nucleophile/leaving group species)—in Tables I and II.[40]

From Tables I and II, it is straightforward to discern that the QM-augmented model outperforms BoB, SLATM, and FCHL19 in combination with KRR for even the smallest number of labeled data points (125 or 225 points). In the absence of QM augmentation, however, the baseline GNN performs significantly worse than any of these methods across the board, as does Chemprop for all but the biggest training/validation set sizes considered. Further underscoring the excellent performance of the ml-QM-GNN model is the lack of hyperparameter optimization and the use of the mean square error loss during training, rather than MAE directly (cf. Sec. S1). Nevertheless, it should also be noted that simple one-hot encoding combined with KRR still outperforms the QM-augmented model here. However, as will be demonstrated below, one-hot encoding—as well as other models that do not base their representation on physical principles—suffers from an inherent limitation related to generalization, i.e., out-of-sample predictions, which severely limits their appeal with respect to our ml-QM-GNN in practical applications.

As an additional check, we constructed multivariate linear models based on the QM descriptor values outputted by our surrogate model, cf. the final columns in Tables I and II, as well as Sec. S8 of the supplementary material. The accuracies of these models are significantly worse than those of the (non-linear) ml-QM-GNN model, and the performance does not increase as more data points are added to the training set. Furthermore, univariate analysis reveals that none of the descriptors correlate particularly well with the activation energies: The maximal R^2 values obtained are 0.24 and 0.44 for the the S_N2 and E2 reactions, respectively. These findings confirm the established insight from the cDFT/VB perspectives that, even though these electronic descriptors determine chemical reactivity to a great extent (*vide infra*), there is no simple, universally valid linear relationship between their magnitude and the height of reaction barriers, i.e., non-linearity is required to fully exploit the reactivity patterns encoded in these physically motivated descriptors.[15] Additional evidence for the latter point can be found in Sec. S13 of the supplementary material: (non-linear) GNNs that start from an exclusively QM-based representation do recover a

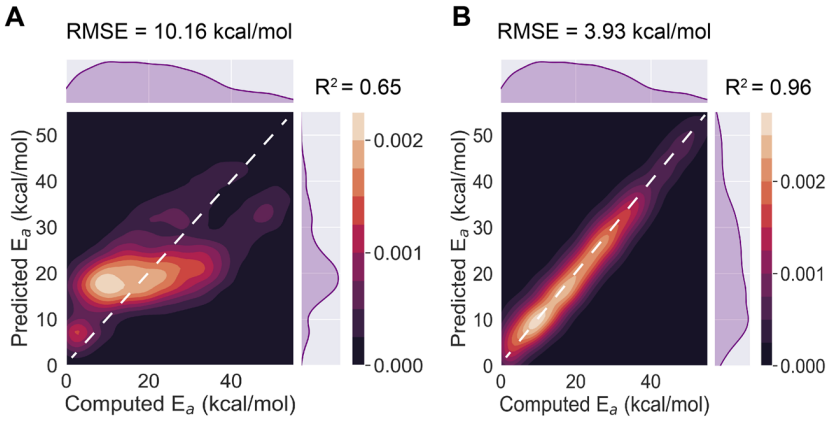

FIG. 3. Correlation plots for (a) the regular GNN model and (b) the ml-QM-GNN model applied to the full E2/S_N2 activation energy dataset from three 5-fold cross-validations in a 60/20/20-split. The colorbars indicate the scale of the 2D kernel density estimate plots. The standard deviations were determined based on the MAEs for the three replicates.

reasonable accuracy, although they still underperform with respect to the full ml-QM-GNNs.

Since our WL-based GNN models involve pooling over reacting atoms, they enable simultaneous treatment of distinct reaction modes for the same reactant/reagent system (in contrast to the reaction-specific KRR models). Hence, we were able to combine the data for the E2 and S_N2 reactions and train a common model for both sets of barrier heights. In Fig. 3, correlation plots between the "true" (computed) and predicted activation energies (E_a), aggregated across all test sets sampled during three 5-fold CVs, are presented.

The correlation between the predicted and computed values is weak for our baseline GNN model [Fig. 3(a); R^2 = 0.65]; the mean MAE and root mean square error (RMSE) from three replicates is 8.4 and 10.1 kcal/mol, respectively. The QM-augmented analog achieves much stronger correlation [Fig. 3(b); R^2 = 0.96] and an average MAE and RMSE of 2.9 and 3.9 kcal/mol.

B. Generalizability—E2/S_N2 regression dataset

Next, we considered the ability of our regression models to generalize to new compounds not present in the training data. While one-hot encoding representations have proven to be well suited for reactivity problems focused on interpolation,[73,74] they do not perform well on in out-of-sample predictions.[49]

To assess model generalizability, training/validation and test sets were selectively sampled so that for one of the four nucleophiles in the dataset, all its reactions would consistently be part of the test set, and consequently, this nucleophile is not "seen" by the model during training. The random train and validation set sampling (in a 3:1 ratio) was iterated five times, and the resulting predictions were aggregated. In Fig. 4, the correlation plots for each of the different "held-out" nucleophiles are presented.

The regular, WL-based GNN models barely manage to reproduce the qualitative trend in the computed activation energies; they significantly underestimate the quantitative values for the barriers when either the hydride or fluoride nucleophiles are held-out, whereas the barriers are vastly underestimated when chloride and bromide are held-out during training (mean RMSEs range from 9 to 18 kcal/mol). KRR in combination with one-hot encoding[40] performs even worse than the regular GNN on this task; upon selectively sampling, RMSEs between 9 and 20 kcal/mol are obtained for this model architecture (cf. Table S5 of the supplementary material). Remarkably, the FCHL19 and SLATM representations do not fare much better than one-hot encoding (mean RMSEs range from 7 to 35 kcal/mol for SLATM and from 5 to 24 kcal/mol for FCHL19, cf. Table S6). Finally, Chemprop also fails to produce reasonable predictions (mean RMSEs between 9 and 16 kcal/mol, cf. Table S11 of the supplementary material), underscoring that limited generalizability is a universal issue for models that do not base their representation on physical principles. The ml-QM-GNN on the other hand obtains decent correlations for each of the models trained with hold-out nucleophiles [Figs. 4(e)–4(h)], and the quantitative agreement between model predictions and true values is much better: For hydride and fluoride, the mean RMSEs amount to 8–9 kcal/mol, whereas for chloride and bromide, the mean RMSEs amount to a reasonable 5–6 kcal/mol.

These results constitute an unequivocal demonstration that designing a GNN model so that it constructs a QM-based representation before the final reactivity prediction not only improves the model accuracy in this data-limited setting but also improves the model's ability to generalize to unseen nucleophiles, i.e., nucleophiles not found in any example in either the training or the validation set.

C. Explainability—E2/S_N2 dataset

Finally, we aimed to gain some insights into how the model reaches its decisions/reactivity predictions. It is not always apparent what exactly deep learning models are learning and how they generalize to new, previously unseen data points, which makes their performance less predictable in prospective settings. We performed a set of ablation experiments, where we controlled the number and type of atom-centered QM descriptors that are used to supplement the structural representation. Specifically, we masked either the nucleophilic and electrophilic Fukui indices or the Hirshfeld partial charges and NMR shielding constants (the effect of inclusion of individual descriptors in the model is concisely discussed in Sec. S17 of the supplementary material).

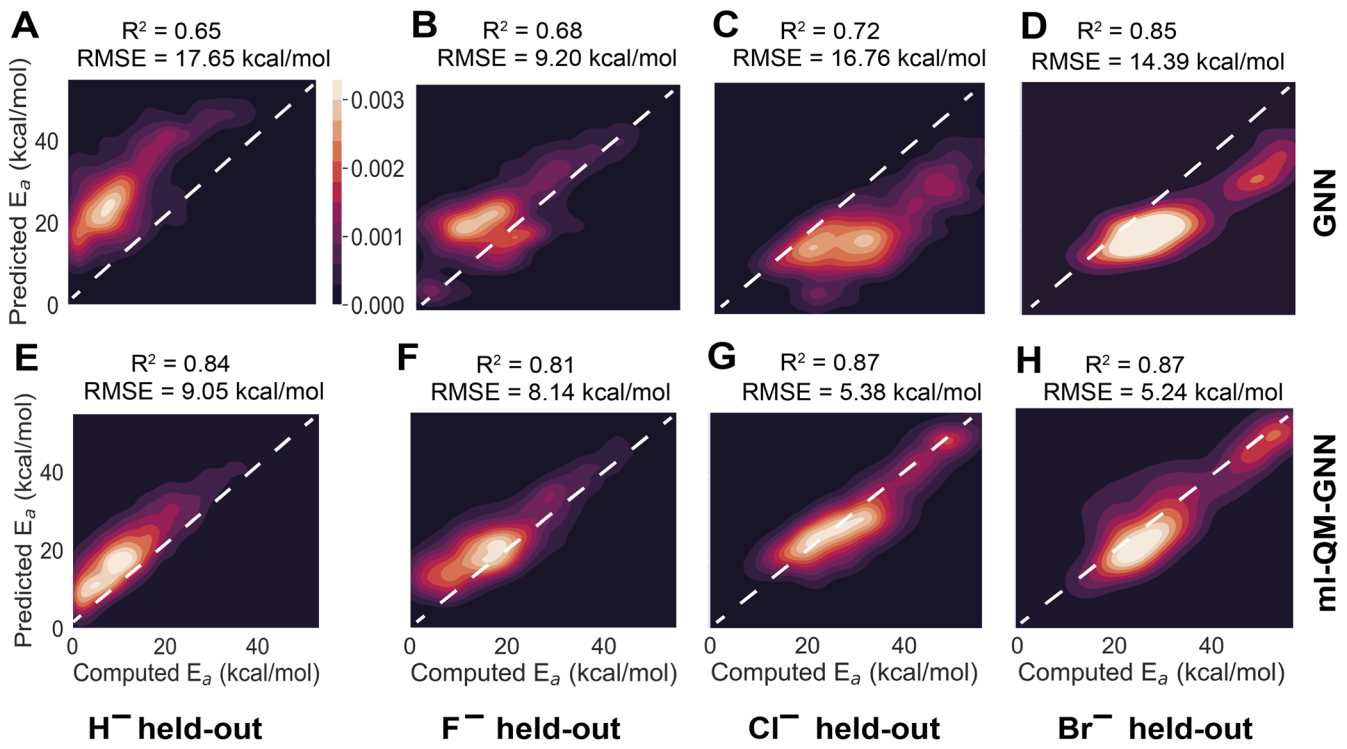

FIG. 4. Correlation plots for the aggregated predictions made by the regular GNN model across iterations, with held-out nucleophiles: (a) H^-, (b) F^-, (c) Cl^-, and (d) Br^-. Correlation plots for the ml-QM-GNN model with held-out nucleophiles: (e) H^-, (f) F^-, (g) Cl^-, and (h) Br^-. The mean RMSE is shown at the top of each individual panel (cf. Sec. S12 for the obtained standard deviations).

The decision to consider the effect of these pairs of descriptors simultaneously is inspired by the core principle emerging from physical organic chemistry/cDFT that the interactions between reacting species can generally be subdivided in two main types: "hard–hard" or electrostatic interactions, and "soft–soft" or (frontier) orbital interactions.[3,75] Fukui functions, defined as the (atom-condensed) distribution of an added and removed electron to the system, probe the latter, whereas atomic charges and NMR shielding constants, which reflect the (de)shielding of the nuclei caused by intramolecular electron donation/withdrawal, probe the former.[5,15,76] It should be noted that this theory-informed dichotomy of the considered atom-condensed descriptors can also readily be retrieved from purely data-driven feature analysis: In both datasets considered in this study, partial charges and NMR shielding constant values correlate with a significant extent, whereas neither of those consistently correlate with either of the Fukui function values, cf. Sec. S18 of the supplementary material.

Table III contains the accuracies obtained after three 5-fold CVs for the baseline GNN model and the models with one or both sets of estimated QM descriptors. For the baseline model, an average RMSE of 10.1 kcal/mol was obtained; for the full QM-augmented model ["ml-QM-GNN (full)"], the average RMSE was reduced to 3.9 kcal/mol. Remarkably, the ablated model that bases its reactivity predictions solely on atomic charges and NMR shielding constants as QM descriptors ["ml-QM-GNN (charge + NMR)"] recovers the exact same accuracy as the full QM-augmented model (RMSE = 3.9 kcal/mol); the ablated model that only makes use of the Fukui functions ["ml-QM-GNN (Fukui)"] on the other hand achieves the same accuracy as the baseline model (RMSE = 10.0 kcal/mol). When the training set size during these fivefold CVs is reduced to 200 data points, similar results are obtained: The ml-QM-GNN (Fukui) model only marginally improves the accuracy of the model relative to the baseline, whereas the ml-QM-GNN (charge + NMR) model gets exceedingly close to the ml-QM-GNN (full) model (cf. Sec. S19). These findings suggest that electrostatic/hard–hard interactions are

TABLE III. Model performance as a function of QM descriptor set inclusion. Mean RMSEs and standard deviations (kcal/mol) obtained for the E2/S_N2 barrier height prediction from three random fivefold cross-validations for the different (ablated) GNN models tested, as well as the corresponding classification accuracies (and their standard deviations) for the prediction of E2 vs S_N2 preference. The standard deviations were again determined from the three replicates.

Model	RMSE (kcal/mol)	Accuracy (%)
Baseline GNN	10.13 ± 0.14	77.0 ± 0.8
ml-QM-GNN (full)	3.92 ± 0.01	89.0 ± 0.6
ml-QM-GNN (Fukui)	10.02 ± 0.02	77.1 ± 0.1
ml-QM-GNN (charge + NMR)	3.93 ± 0.02	88.9 ± 0.6

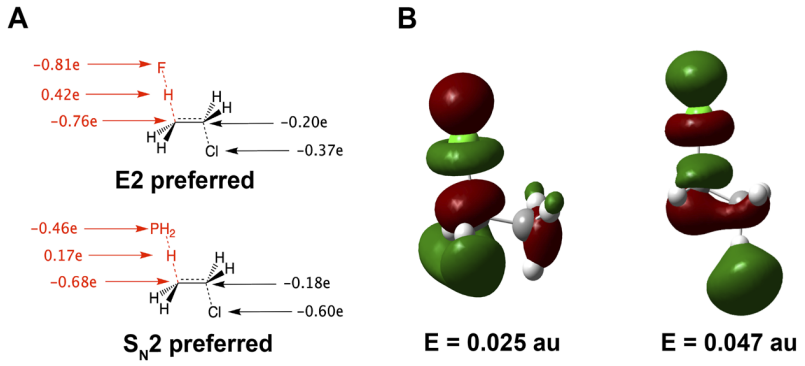

FIG. 5. (a) Partial (NPA) charge distribution in the E2-TS geometry for $F^- + H_3CCH_2Cl$ (top) and $H_2P^- + H_3CCH_2Cl$ (bottom); two model systems frequently used as a starting point for qualitative analyses of E2/S_N2 competition. The more pronounced (−)–(+)–(−) charge array in the case of the F^- nucleophile causes sufficient electrostatic stabilization for the E2-TS to drop below the S_N2-TS in energy; in the case of the H_2P^- nucleophile, the reduced electrostatic stabilization causes the S_N2-TS to remain lower in energy. (b) Lowest unoccupied molecular orbital (LUMO; E = 0.025 a.u.; left) and LUMO + 1 (E = 0.047 a.u.; right) for H_3CCH_2Cl. The calculations were performed at the M06/def2-TZVP level-of-theory (cf. Sec. S20).[77]

the main drivers of the variations observed in the dataset and that Fukui functions, i.e., frontier orbital interactions, are not particularly relevant in this regard.

Further evidence that Fukui functions do not play a significant role in the decision-making process of our QM-augmented network is obtained when the fully trained ml-QM-GNN model is applied to all data points after averaging Fukui function values across all atoms. In this setup, the model predictions are barely affected (R^2 = 0.96 between the original predictions and the predictions with Fukui information averaged out). Averaging out the charges and NMR shielding constants on the other hand scrambles the predictions made by the model entirely (Fig. S6).

It is worthwhile to note here that our uni-/multivariate analysis (*vide supra*) already contained clues hinting in the same direction: While—as noted before—none of the descriptors were found to correlate particularly well with the activation energies in these linear models, the hard–hard descriptors did collectively outperform the soft–soft descriptors by a significant margin (cf. Sec. S8 of the supplementary material).

The same trends discussed above emerge for the classification models constructed for the second curated dataset, also extracted from the E2/S_N2 reaction data (Table III). Again, an improved accuracy is observed when the model bases its predictions on the full set of predicted QM descriptors (77%–89% accuracy; cf. Sec. S15 for some representative failures to predict the correct preference for the ml-QM-GNN model). Removing the Fukui function information from the QM-augmented model does not affect the accuracy in a meaningful way, whereas removing the charge and NMR shielding constant data again causes the accuracy to drop to the baseline level.

Our findings about the relative importance of electrostatic/"hard–hard" vs (frontier) orbital/"soft–soft" interactions, emerging in both regression and classification tasks, are perfectly in line with a recent qualitative VB/cDFT analysis undertaken by one of the authors of the present work.[77] In this analysis, it was demonstrated that the modulation of the E2/S_N2 competition is primarily driven by the electrostatic interactions present in the E2-TS: the formation of a strongly stabilizing array of point-charges, i.e., (−)–(+)–(−), in this geometry tends to push its energy below that of the S_N2-TS; in the case that the point-charges in this array—and thus the Coulombic interaction—are weaker, the S_N2-pathway dominates [cf. Fig. 5(a)].

The observation that the Fukui function values are not really helpful to a GNN aiming to learn this mechanistic competition could also have been readily anticipated from the analysis presented in the same qualitative study: The ethylhalide substrates on which the nucleophiles attack generally carry two relatively close-lying, unoccupied frontier orbitals that are delocalized over both the α-carbon and the hydrogen on the β-carbon [Fig. 5(b)];[77] providing information about only the lowest-lying of these orbitals through the (electrophilic) Fukui function is not very informative.

The discussion above demonstrates that combining a qualitative analysis rooted in conceptual reactivity frameworks with our QM-augmented machine learning approach leads to a productive synergy for the considered dataset of competing E2/S_N2 reactions: On the one hand, the qualitative insights provide context to explain and understand the decision/prediction-making process of the network. At the same time, the results emerging from our ablation study can also be considered as an indirect, data-driven confirmation of the qualitative reactivity analysis.

D. Explaining the predicted regioselectivity of aromatic substitution reactions

To further explore the explainability/interpretability of ml-QM-GNN models, we revisited the classification dataset containing regiochemical data for electrophilic substitution reactions, originally compiled by Guan *et al.* to demonstrate the potential of this model architecture in data-limited settings.[50] As mentioned in Sec. II, we select only those data points that are publicly available in the USPTO database, resulting in a curated dataset of 3242 reactions. In Table IV, the accuracies achieved by the baseline GNN, the ml-QM-GNN, and the two ablated models on the curated dataset (training set limited to 200 data points) are presented. Once more, we observe

TABLE IV. Average accuracies (and their standard deviation across replicates) obtained from three random fivefold CVs for the different (ablated) WL-based GNN models applied to the curated aromatic substitution dataset (training set limited to 200 data points).

Model	Accuracy (%)
Baseline GNN	71.4 ± 1.3
ml-QM-GNN (full)	86.5 ± 0.7
ml-QM-GNN (Fukui)	84.4 ± 1.1
ml-QM-GNN (charge + NMR)	82.2 ± 1.3

that the full QM-augmented model significantly outperforms the baseline GNN. More interestingly, it can be observed now that *each* set of descriptors leads to a significant improvement in accuracy over the baseline, but *both* are needed to maximize performance.

Even though the differences are rather small, the results presented in Table IV suggest that for this class of reactions, the Fukui functions are slightly more informative than charges and NMR shielding constants, which is in contrast to what was observed for the E2/S_N2 dataset in Sec. III C. These findings can be straightforwardly reconciled with the qualitative physical organic reactivity models that have been constructed and popularized throughout the years: Aromatic compounds are strongly delocalized, and hence, orbital interactions/changes in delocalization stabilization as probed through the Fukui functions are generally considered to be the main driving force shaping their reactivity.[14,78,79] At the same time, it has been underscored in recent years that electrostatics cannot be neglected—in particular under conditions favoring kinetic control—since atomic charges tend to become more pronounced as the compounds involved in the reaction approach,[9] enabling a significant Coulombic stabilization/destabilization of the wave function in the transition state region.[15,76]

To obtain a better understanding of how the (ablated) QM-augmented models reach their decisions/predictions, we constructed a set of confusion matrices comparing the respective prediction accuracies for all test sets considered during the first cross-validation (Fig. 6).

These matrices show that there are relatively few data points (97 and 119) where an ablated model makes the correct prediction, while the model with access to all descriptors reaches the wrong conclusion; the full QM-augmented model appears to mainly rectify incorrect predictions by the ablated models (cf. the upper two square matrices in Fig. 6). Additionally, the lower-right confusion matrix in Fig. 6 demonstrates that many incorrect predictions by the two ablated models are distinct: The number of points for which there is disagreement between the models (245 + 351) is greater than the off-diagonal elements in the upper matrices.

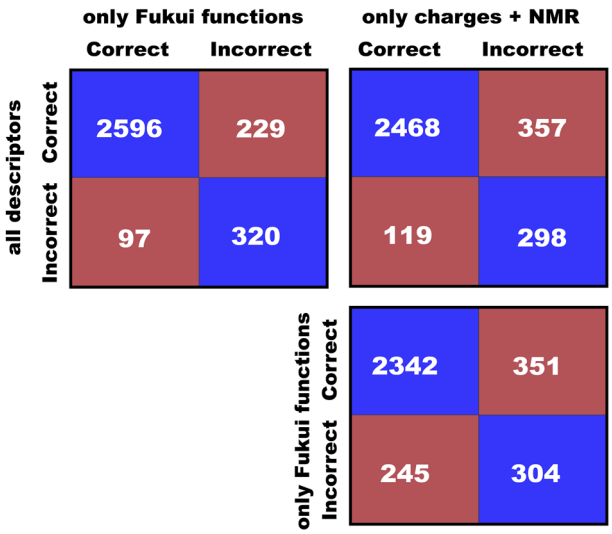

FIG. 6. Confusion matrices comparing the predictions made by the different (ablated) ml-QM-GNN models for all test sets sampled during the first fivefold CV combined. The labels in the margins of the individual matrices indicate which descriptors are considered by the respective model, i.e., either only the soft–soft or hard–hard descriptors or both descriptor-types combined.

We then considered whether incorrect prediction of regioselectivity preference is connected to failures of either the "hard–hard"/electrostatic or "soft–soft"/(frontier) orbital criterion. The "hard–hard"/electrostatic criterion is considered to be fulfilled when the reacting site on the aromatic substrate corresponds to the site carrying the highest partial negative charge (in the case of electrophilic attack) or partial positive charge (in the case of nucleophilic attack) since this would maximize the Coulombic stabilization upon approach between the reacting species. The "soft–soft"/(frontier) orbital criterion is considered to be fulfilled when the reacting site on the substrate corresponds to the site on which the nucleophilic Fukui function is most concentrated (in the case of electrophilic attack) or the site on which the electrophilic Fukui function is most concentrated (in the case of nucleophilic attack)—which would maximize the orbital interaction upon approach between the reacting species.

Since nitration reactions and halogenation reactions involving *N*-substituted succinimides are electrophilic without exception, are straightforward to recognize based on a SMILES representation, and collectively cover approximately half of the dataset (1534 out of 3242 data points), we decided to focus exclusively on these reaction types for this part of our analysis. A confusion matrix comparing the predictions obtained through naive evaluation of the individual physical organic criteria (*vide supra*) with the predictions made by the (ablated) models during fivefold CV is presented in Table S14. The predictions made by the models tend to adhere to the respective criteria: The vast majority of data points for which both criteria point to the correct reactive site are classified correctly by the ablated and the full QM-augmented GNNs, whereas reactions incorrectly classified only by the ablated model that exclusively considers Fukui function values disproportionately violate the "soft–soft" criterion (and vice versa: The reactions that are only incorrectly classified by the ml-QM-GNN (charge + NMR) disproportionately violate the "hard–hard" criterion, Fig. 7). Only a handful of reactions for which neither criterion is fulfilled are classified correctly.

Putting all this together, one can conclude that the trained QM-augmented GNN model with all descriptors included appears to be able to balance the relative importance of the electrostatic and (frontier) orbital criteria, whereas the ablated models are generally biased toward (a) assigning too much importance to the interaction type that they have access to and (b) adhering to the rudimentary assignment based on the "soft–soft"/"hard–hard" criterion, even when this assignment is incorrect.

As such, our ml-QM-GNN model essentially refines the balancing act that is implicitly part of any cDFT-based qualitative reactivity analysis. As indicated above, most human theoreticians tend to assume that the regioselectivity of aromatic reactions is primarily determined by the Fukui functions, i.e., the soft–soft interactions dominate, and they will often ignore the electrostatics altogether (or assume that these are in sync with the Fukui function values).[79] Our analysis (cf. Sec. S21) shows that this is not an unreasonable approximation to make for this dataset; selecting the site to

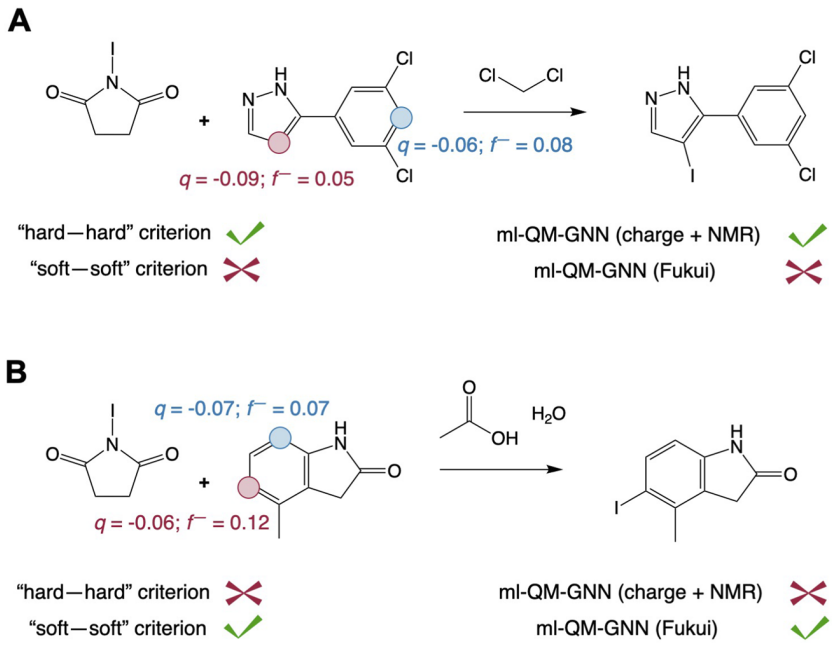

FIG. 7. The predictions made by the ablated models generally adhere to their corresponding physical organic criterion (q denotes the predicted partial charge, and f^- denotes the predicted atom-condensed nucleophilic Fukui function value; the red dots correspond to the reported/"true" reactive sites, and the blue dots are competing sites). (a) For this example reaction, the hard–hard criterion is fulfilled, but the soft–soft one is not; correspondingly, the ablated model that considers exclusively the partial charges an NMR shielding constants makes the correct prediction and the ablated model that considers exclusively Fukui function values makes the incorrect one. (b) Example reaction for which the situation is reversed.

which the soft–soft criterion points—ignoring the hard–hard criterion completely—results in an accuracy of ≈80%. By relaxing this rigid guiding rule and replacing it by a more complex classification function centered around the same QM descriptors, our models already reach accuracies exceeding 86% when a mere 200 training points are used (Table IV) and reaching up to 93%–94% when the training set size is expanded to 1945 training points (cf. Sec. S6 of the supplementary material for learning curves visualizing the gradual rise in accuracy upon training set expansion).

E. Failures in predicting regioselectivity of aromatic substitution reactions

Finally, we also took a closer look at some of the reacting systems for which neither criterion is fulfilled and for which all our GNN models, trained on only 200 example reactions, fail to make the correct prediction. While it appears impossible to assign a specific reason for each failure, there are some recurring patterns.

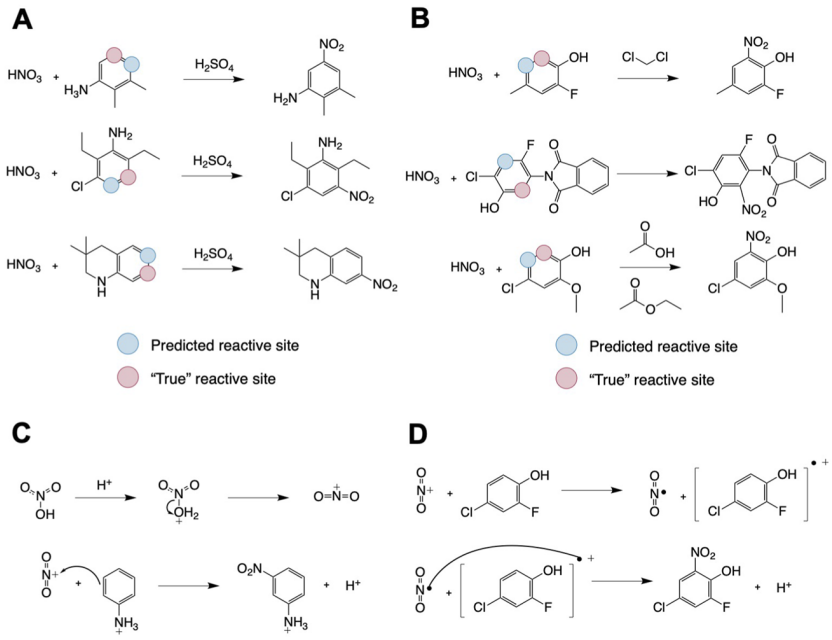

FIG. 8. (a) Some examples of nitration reactions with an NH_2-substituent on the aromatic substrate for which both physical organic criteria and all QM-augmented models fail (the blue dots indicate the—incorrectly—predicted sites; the red dots indicate the true reactive sites). (b) Some examples of nitration reactions with an OH-substituent on the aromatic substrate for which both physical organic criteria and all QM-augmented models fail. (c) Mechanism of a regular (electrophilic) nitration reaction involving a (protonated) aniline substrate. (d) Proposed radical mechanism for nitration of halogenated phenol analogs.[80]

As an example, all criteria and models consistently appear to fail for most nitration reactions involving aniline analogs, i.e., aromatic substrates with an NH_2-substituent [Fig. 8(a)]. For this type of reaction, nitration in ortho- or para-position is expected, but the "true" reactive site corresponds to the meta-position. This failure is a reflection of the fact that nitration reactions are usually performed in strongly acidic reaction media to promote the protonation of nitric acid into the active nitronium ion species, which inherently results in simultaneous protonation of the NH_2-substituent, rendering this substituent meta-directing instead of para-directing [Fig. 8(c)].

Another recurring failure is observed for nitration reactions involving halogenated phenols. In this type of reaction, both the criteria and models predict the reaction to occur in ortho-/para-position relative to the halogen substituent, whereas the "true"/recorded reactive site is consistently the ortho-position relative to the OH-substituent [Fig. 8(b)]. This failure can most likely be attributed to a mechanistic crossover: Instead of a conventional electrophilic aromatic substitution mechanism, reactions between phenolic compounds and nitric acid/nitronium ions have been demonstrated to involve a (pure) single electron transfer step, followed by radical recombination [Fig. 8(d)].[80,81] This change in lthe mechanism can be expected to impact the balance in directing strengths of the halogen- and hydroxy-substituents with respect to regular electrophilic aromatic substitution reactions, resulting in a modification of the regiochemistry.

Since both the physical organic criteria and our GNN models are agnostic to reaction conditions and "concealed" reaction steps, they are unable to capture these modifications to the directing character of the NH_2 and OH-substituents under data-scarce conditions. One can anticipate that by adding more and more training data, the model may learn to overrule the "regular" QM descriptor assignment when NH_2/OH-substituents and nitric acid are simultaneously present in the reacting system, but since only 200 data points were used for the initial training here, there are simply not enough examples to capture these particular patterns. Indeed, we find qualitative evidence for this assertion in the case of the nitration reactions of halogenated phenols: With 1945 training points, the correct regiochemical predictions are generally recovered by the ml-QM-GNN model for this class of failures (cf. Sec. S22 of the supplementary material).

IV. CONCLUSIONS

In this work, the performance, explainability, and generalizability of the QM-augmented GNN (ml-QM-GNN) model architecture have been assessed for a few distinct predictive chemistry tasks. Our models achieve a significantly improved accuracy over an analogous, conventional GNN baseline and generalize markedly better to unseen compounds, particularly in a data-limited regime. Even when only a couple hundred labeled data points are available, our ml-QM-GNN models are competitive with traditional "low-data" KRR models.

Importantly, since the predictions made by our models are rooted in (predicted) QM descriptors, it becomes possible to build bridges between their predictions and the existing physical organic frameworks, developed to qualitatively analyze chemical reactivity. Through a series of selective descriptor ablation experiments, we have demonstrated that for competing E2/S_N2 reactions, Fukui function values, i.e., information about soft–soft interactions, are of limited value; charges and NMR shielding constants are the main drivers of the improvement in model accuracy with respect to the baseline GNN model. These findings can be rationalized through consideration of a recent qualitative valence bond/cDFT analysis of this mechanistic competition.[77]

For aromatic substitution reactions, we observed that the ml-QM-GNN model appears to make its decision/predictions in a similar manner as human theoreticians trained in physical organic chemistry/cDFT, i.e., by considering soft–soft and hard–hard interactions separately through their corresponding local descriptors. What makes our models excel with respect to a naive conceptual treatment is their ability to fine-tune the relative importance of the individual physical organic criteria based on subtle patterns present in the data, resulting in a more complex classification function and a significantly higher accuracy.

Overall, our analysis underscores that a productive interplay between machine learning models and qualitative reactivity analysis is possible: On the one hand, qualitative insights into the considered reactivity problem provide context to explain and understand the decision-making process of the network. Additionally, they can provide clues about the suitability to include specific QM descriptors in the neural network. At the same time, the results emerging from machine learning models augmented with QM descriptor information can provide an indirect, data-driven confirmation of a qualitative reactivity analysis.

SUPPLEMENTARY MATERIAL

See the supplementary material for an in-depth technical description of the WL-based GNN model architectures, technical details related to the overall training process and cross-validation, summarizing statistics to characterize the data distributions, the effect of removing "duplicate" data points from the E2/S_N2 dataset, note regarding the data curation procedure for the E2/S_N2 classification dataset, learning curves for the aromatic substitution dataset, computational details related to the application of Chemprop to the E2/S_N2 dataset, multi- and univariate analysis for the E2/S_N2 dataset, selective sampling results for one-hot encoding in combination with KRR, selective sampling results for Chemprop, RMSEs and standard deviations obtained for the WL-based GNN and ml-QM-GNN during selective sampling, performance of a GNN with an exclusively QM-based representation as input for the E2/S_N2 dataset, the effect of averaging out the respective descriptors for the full QM-augmented model trained on the E2/S_N2 dataset, representative examples of failures of the ml-QM-GNN model for the E2/S_N2 classification, performance of the full ml-QM-GNN model, trained exclusively on either the E2 or S_N2 data, on the held-out reaction type, the effect of inclusion of individual QM descriptors in the ml-QM-GNN, correlation analysis of the QM descriptors included in the ml-QM-GNN, methodology for the DFT calculations in Fig. 5, confusion matrix comparing the classifications by the QM-augmented models to the physical organic criteria for the aromatic substitution dataset, and recovery of the correct regiochemical predictions as the training set increases for the aromatic substitution dataset.

ACKNOWLEDGMENTS

The authors acknowledge the Machine Learning for Pharmaceutical Discovery and Synthesis (MLPDS) consortium for funding and thank both Yanfei Guan and Esther Heid for helpful discussions.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are openly available in the GitHub repository containing the main coding associated with this project at https://github.com/coleygroup/QM-augmented_GNN.

REFERENCES

[1] R. B. Woodward and R. Hoffmann, "Stereochemistry of electrocyclic reactions," J. Am. Chem. Soc. **87**, 395–397 (1965).
[2] M. G. Evans and M. Polanyi, "Further considerations on the thermodynamics of chemical equilibria and reaction rates," Trans. Faraday Soc. **32**, 1333–1360 (1936).
[3] R. G. Pearson, "Hard and soft acids and bases," J. Am. Chem. Soc. **85**, 3533–3539 (1963).
[4] R. G. Parr and R. G. Pearson, "Absolute hardness: Companion parameter to absolute electronegativity," J. Am. Chem. Soc. **105**, 7512–7516 (1983).
[5] P. Geerlings, F. De Proft, and W. Langenaeker, "Conceptual density functional theory," Chem. Rev. **103**, 1793–1874 (2003).
[6] I. Fernández and F. M. Bickelhaupt, "The activation strain model and molecular orbital theory: Understanding and designing chemical reactions," Chem. Soc. Rev. **43**, 4953–4967 (2014).
[7] F. M. Bickelhaupt and K. N. Houk, "Analyzing reaction rates with the distortion/interaction-activation strain model," Angew. Chem., Int. Ed. **56**, 10070–10086 (2017).
[8] S. S. Shaik, "What happens to molecules as they react? A valence bond approach to reactivity," J. Am. Chem. Soc. **103**, 3692–3701 (1981).
[9] S. Shaik and A. Shurki, "Valence bond diagrams and chemical reactivity," Angew. Chem., Int. Ed. **38**, 586–625 (1999).
[10] R. G. Parr and W. Yang, "Density functional approach to the frontier-electron theory of chemical reactivity," J. Am. Chem. Soc. **106**, 4049–4050 (1984).
[11] S. S. Shaik and P. C. Hiberty, *A Chemist's Guide to Valence Bond Theory* (John Wiley & Sons, 2007).
[12] R. Hoffmann, S. Shaik, and P. C. Hiberty, "A conversation on VB vs MO theory: A never-ending rivalry?," Acc. Chem. Res. **36**, 750–756 (2003).
[13] W. T. Borden, R. Hoffmann, T. Stuyver, and B. Chen, "Dioxygen: What makes this triplet diradical kinetically persistent?," J. Am. Chem. Soc. **139**, 9010–9018 (2017).
[14] T. Stuyver, F. De Proft, P. Geerlings, and S. Shaik, "How do local reactivity descriptors shape the potential energy surface associated with chemical reactions? The valence bond delocalization perspective," J. Am. Chem. Soc. **142**, 10102–10113 (2020).
[15] T. Stuyver and S. Shaik, "Unifying conceptual density functional and valence bond theory: The hardness–softness conundrum associated with protonation reactions and uncovering complementary reactivity modes," J. Am. Chem. Soc. **142**, 20002–20013 (2020).
[16] J. N. Wei, D. Duvenaud, and A. Aspuru-Guzik, "Neural networks for the prediction of organic chemistry reactions," ACS Cent. Sci. **2**, 725–732 (2016).
[17] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, and K. F. Jensen, "Prediction of organic reaction outcomes using machine learning," ACS Cent. Sci. **3**, 434–443 (2017).
[18] V. H. Nair, P. Schwaller, and T. Laino, "Data-driven chemical reaction prediction and retrosynthesis," Chimia **73**, 997–1000 (2019).
[19] A. M. Żurański, J. I. Martinez Alvarado, B. J. Shields, and A. G. Doyle, "Predicting reaction yields via supervised learning," Acc. Chem. Res. **54**, 1856–1865 (2021).
[20] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, and A. G. Doyle, "Predicting reaction performance in C–N cross-coupling using machine learning," Science **360**, 186–190 (2018).
[21] P. Schwaller, A. C. Vaucher, T. Laino, and J.-L. Reymond, "Prediction of chemical reaction yields using deep learning," Mach. Learn.: Sci. Technol. **2**, 015016 (2021).
[22] P. C. St. John, Y. Guan, Y. Kim, S. Kim, and R. S. Paton, "Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost," Nat. Commun. **11**, 2328 (2020).
[23] M. Wen, S. M. Blau, E. W. C. Spotte-Smith, S. Dwaraknath, and K. A. Persson, "BonDNet: A graph neural network for the prediction of bond dissociation energies for charged molecules," Chem. Sci. **12**, 1858–1868 (2021).
[24] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems*, Vol. 28, edited by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Curran Associates, Inc., 2015).
[25] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, and K. F. Jensen, "A graph-convolutional neural network model for the prediction of chemical reactivity," Chem. Sci. **10**, 370–377 (2019).
[26] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," J. Chem. Inf. Model. **50**, 742–754 (2010).
[27] F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, and F. Glorius, "A structure-based platform for predicting chemical reactivity," Chem **6**, 1379–1390 (2020).
[28] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction," ACS Cent. Sci. **5**, 1572–1583 (2019).
[29] Q. Yang, V. Sresht, P. Bolgar, X. Hou, J. L. Klug-McLeod, C. R. Butler *et al.*, "Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space," Chem. Commun. **55**, 12152–12155 (2019).
[30] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016), http://www.deeplearningbook.org.
[31] R. Marcinkevičs and J. E. Vogt, "Interpretability and explainability: A machine learning zoo mini-tour," arXiv:2012.01805 (2020).
[32] P. Polishchuk, "Interpretation of quantitative structure–activity relationship models: Past, present, and future," J. Chem. Inf. Model. **57**, 2618–2639 (2017).
[33] Z. L. Niemeyer, A. Milo, D. P. Hickey, and M. S. Sigman, "Parameterization of phosphine ligands reveals mechanistic pathways and predicts reaction outcomes," Nat. Chem. **8**, 610–617 (2016).
[34] Y. Amar, A. M. Schweidtmann, P. Deutsch, L. Cao, and A. Lapkin, "Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis," Chem. Sci. **10**, 6697–6706 (2019).
[35] K. Wu and A. G. Doyle, "Parameterization of phosphine ligands demonstrates enhancement of nickel catalysis via remote steric effects," Nat. Chem. **9**, 779–784 (2017).
[36] A. Pappu and B. Paige, "Making graph neural networks worth it for low-data molecular machine learning," arXiv:2011.12203 (2020).
[37] P. Friederich, G. dos Passos Gomes, R. De Bin, A. Aspuru-Guzik, and D. Balcells, "Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex," Chem. Sci. **11**, 4584–4601 (2020).
[38] L. C. Gallegos, G. Luchini, P. C. St. John, S. Kim, and R. S. Paton, "Importance of engineered and learned molecular representations in predicting organic reactivity, selectivity, and chemical properties," Acc. Chem. Res. **54**, 827–836 (2021).
[39] B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld, and C. Corminboeuf, "Machine learning meets volcano plots: Computational discovery of cross-coupling catalysts," Chem. Sci. **9**, 7069–7077 (2018).
[40] S. Heinen, G. F. von Rudorff, and O. A. von Lilienfeld, "Toward the design of chemical reactions: Machine learning barriers of competing mechanisms in reactant space," J. Chem. Phys. **155**, 064105 (2021).

[41] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space," J. Phys. Chem. Lett. **6**, 2326–2331 (2015).
[42] A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole von Lilienfeld, "FCHL revisited: Faster and more accurate quantum machine learning," J. Chem. Phys. **152**, 044107 (2020).
[43] B. M. Axilrod and E. Teller, "Interaction of the van der Waals type between three atoms," J. Chem. Phys. **11**, 299–300 (1943).
[44] M. S. Sigman, K. C. Harper, E. N. Bess, and A. Milo, "The development of multidimensional analysis tools for asymmetric catalysis and beyond," Acc. Chem. Res. **49**, 1292–1301 (2016).
[45] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, and S. E. Denmark, "Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning," Science **363**, eaau5631 (2019).
[46] W. Beker, E. P. Gajewska, T. Badowski, and B. A. Grzybowski, "Prediction of major regio-, site-, and diastereoisomers in Diels–Alder reactions by using machine-learning: The importance of physically meaningful descriptors," Angew. Chem., Int. Ed. **58**, 4515–4519 (2019).
[47] X. Li, S. Q. Zhang, L. C. Xu, and X. Hong, "Predicting regioselectivity in radical C–H functionalization of heterocycles through machine learning," Angew. Chem., Int. Ed. **59**, 13253–13259 (2020).
[48] K. Jorner, T. Brinck, P.-O. Norrby, and D. Buttar, "Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies," Chem. Sci. **12**, 1163–1175 (2021).
[49] J. G. Estrada, D. T. Ahneman, R. P. Sheridan, S. D. Dreher, and A. G. Doyle, "Response to Comment on 'Predicting reaction performance in C–N cross-coupling using machine learning,'" Science **362**, eaat8763 (2018).
[50] Y. Guan, C. W. Coley, H. Wu, D. Ranasinghe, E. Heid, T. J. Struble, L. Pattanaik, W. H. Green, and K. F. Jensen, "Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors," Chem. Sci. **12**, 2198–2208 (2021).
[51] G. F. von Rudorff, S. N. Heinen, M. Bragato, and O. A. von Lilienfeld, "Thousands of reactants and transition states for competing E2 and S2 reactions," Machine Learn: Sci. Technol. **1**, 045026 (2020).
[52] G. Landrum *et al.*, "RDKit: Open-source cheminformatics," http://www.rdkit.org (2006).
[53] Z. Liu, L. Lin, Q. Jia, Z. Cheng, Y. Jiang, Y. Guo, and J. Ma, "Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties via multitask learning," J. Chem. Inf. Model. **61**, 1066–1082 (2021).
[54] R. Zubatyuk, J. S. Smith, J. Leszczynski, and O. Isayev, "Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network," Sci. Adv. **5**, eaav6490 (2019).
[55] E. D. Glendening, C. R. Landis, and F. Weinhold, "NBO 6.0: Natural bond orbital analysis program," J. Comput. Chem. **34**, 1429–1437 (2013).
[56] F. L. Hirshfeld, "Bonded-atom fragments for describing molecular charge densities," Theor. Chim. Acta **44**, 129–138 (1977).
[57] K. Wolinski, J. F. Hinton, and P. Pulay, "Efficient implementation of the gauge-independent atomic orbital method for NMR chemical shift calculations," J. Am. Chem. Soc. **112**, 8251–8260 (1990).
[58] W. Yang and W. J. Mortier, "The use of global and local molecular parameters for the analysis of the gas-phase basicity of amines," J. Am. Chem. Soc. **108**, 5708–5711 (1986).
[59] W. Jin, C. Coley, R. Barzilay, and T. Jaakkola, "Predicting organic reaction outcomes with Weisfeiler-Lehman network," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017), Vol. 30.
[60] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, 2016), pp. 1480–1489.
[61] R. A. Kendall, T. H. Dunning, Jr., and R. J. Harrison, "Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions," J. Chem. Phys. **96**, 6796–6806 (1992).
[62] C. Hampel, K. A. Peterson, and H.-J. Werner, "A comparison of the efficiency and accuracy of the quadratic configuration interaction (QCISD), coupled cluster (CCSD), and Brueckner coupled cluster (BCCD) methods," Chem. Phys. Lett. **190**, 1–12 (1992).
[63] M. J. Frisch, J. A. Pople, and J. S. Binkley, "Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets," J. Chem. Phys. **80**, 3265–3269 (1984).
[64] M. Schütz and F. R. Manby, "Linear scaling local coupled cluster theory with density fitting. Part I: 4-external integrals," Phys. Chem. Chem. Phys. **5**, 3349–3358 (2003).
[65] A. D. McLean and G. S. Chandler, "Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, $Z = 11–18$," J. Chem. Phys. **72**, 5639–5648 (1980).
[66] R. Krishnan, J. S. Binkley, R. Seeger, and J. A. Pople, "Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions," J. Chem. Phys. **72**, 650–654 (1980).
[67] T. H. Dunning, Jr., "Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen," J. Chem. Phys. **90**, 1007–1023 (1989).
[68] J. Jensen, xyz2mol, GitHub repository, 2020.
[69] C. W. Coley, W. H. Green, and K. F. Jensen, "RDChiral: An RDKit wrapper for handling stereochemistry in retrosynthetic template extraction and application," J. Chem. Inf. Model. **59**, 2529–2537 (2019).
[70] N. Software, Pistachio, 2021.
[71] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea *et al.*, "Analyzing learned molecular representations for property prediction," J. Chem. Inf. Model. **59**, 3370–3388 (2019).
[72] H. Dai, B. Dai, and L. Song, "Discriminative embeddings of latent variable models for structured data," in *International Conference on Machine Learning (PMLR)* (International Machine Learning, Society, 2016), pp. 2702–2711.
[73] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, and L. Cronin, "Controlling an organic synthesis robot with machine learning to search for new reactivity," Nature **559**, 377–381 (2018).
[74] K. V. Chuang and M. J. Keiser, "Comment on 'Predicting reaction performance in C–N cross-coupling using machine learning,'" Science **362**, aat8603 (2018).
[75] G. Klopman, "Chemical reactivity and the concept of charge-and frontier-controlled reactions," J. Am. Chem. Soc. **90**, 223–234 (1968).
[76] J. S. M. Anderson, J. Melin, and P. W. Ayers, "Conceptual density-functional theory for general chemical reactions, including those that are neither charge- nor frontier-orbital-controlled. 1. Theory and derivation of a general-purpose reactivity indicator," J. Chem. Theory Comput. **3**, 358–374 (2007).
[77] T. Stuyver and S. Shaik, "Resolving entangled reactivity modes through external electric fields and substitution: Application to E_2/S_N2 reactions," J. Org. Chem. **86**, 9030–9039 (2021).
[78] K. Fukui, T. Yonezawa, C. Nagata, and H. Shingu, "Molecular orbital theory of orientation in aromatic, heteroaromatic, and other conjugated molecules," J. Chem. Phys. **22**, 1433–1442 (1954).
[79] W. Langenaeker, K. Demel, and P. Geerlings, "Quantum-chemical study of the Fukui function as a reactivity index: Part 2. Electrophilic substitution on mono-substituted benzenes," J. Mol. Struct.: THEOCHEM **234**, 329–342 (1991).
[80] L. Ducry and D. M. Roberge, "Controlled autocatalytic nitration of phenol in a microreactor," Angew. Chem., Int. Ed. **44**, 7972–7975 (2005).
[81] C. L. Perrin, "Necessity of electron transfer and a radical pair in the nitration of reactive aromatics," J. Am. Chem. Soc. **99**, 5516–5518 (1977).